ON THE ADVERSARIAL ROBUSTNESS OF SUBSPACE LEARNING

Fuwei Li, Lifeng Lai, and Shuguang Cui

Department of ECE, University of California, Davis
Email:{fli,lflai,sgcui}@ucdavis.edu

ABSTRACT

In this paper, we investigate the adversarial robustness of subspace learning problems. Different from the scenario addressed by classic robust algorithms that assume fractions of data are corrupted, we consider a more powerful adversary who can observe the whole data and modify all of them. The goal of the adversary is to maximize the distance between the subspace learned from the original data set and that learned from the modified data. We characterize the optimal rank-one attack strategy and show that the optimal strategy depends on the smallest singular value of the original data matrix and the adversary's energy budget.

***Index Terms*—** Principal component analysis, subspace learning, adversarial robustness

1. INTRODUCTION

Subspace learning has applications in many areas such as surveillance video analysis, recommendation system , etc [1, 2, 3]. Furthermore, many interesting work have proposed robust subspace learning algorithms that are capable of mitigating the impact of random noise or certain percentage of outliers presented in the data set [4].

Motivated by the fact that machine learning algorithms are being increasingly used in safety critical applications and security related applications [5, 6, 7], we investigate the adversarial robustness of subspace learning algorithms. In particular, we examine the robustness of subspace learning algorithms against not only random noise or unintentional corrupted data, but also malicious data produced by powerful adversaries who can modify the whole data set. Our study is related to the growing list of interesting papers on adversarial machine learning. These papers have revealed that many learning algorithms are vulnerable to adversaries [8, 9, 10, 11]. One notable instance is the adversarial example phenomenon [12, 13, 14] in deep learning where an adversary can introduce small but carefully crafted changes into the images so as to mislead the neural network to make incorrect decisions.

The work of F. Li and S. Cui was supported in part by grant NSFC-61629101, by NSF with grants DMS-1622433, AST-1547436, ECCS-1659025. The work of L. Lai was supported by National Science Foundation under Grants CCF-1717943 and ECCS-1711468.

In our problem formulation, given the original data matrix from which we will learn a low dimension subspace via principal component analysis (PCA), a powerful adversary can modify all entries of this data matrix. The goal of the adversary is to maximize the distance between the subspace learned from the original data matrix and the subspace learned from the modified data matrix. In this paper, we use Asimov distance to measure the distance between subspaces. We assume that the adversary can use the whole data set to carefully construct a rank-one attack matrix and add it to the original data set. We characterize the optimal attack strategy in terms of the adversary's energy budget and its impact on the learned subspace. We show that the optimal adversarial strategy depends only on the smallest singular value of the original data matrix. If the energy budget is larger than the smallest singular value, the attacker can construct an attack matrix to make the Asimov distance to be $\pi/2$, the largest possible value. If the energy budget is less than the smallest singular value, we show that the optimal attack matrix must adopt a certain form and the corresponding Asimov distance is directly related to the ratio between the energy budget and the smallest singular value. Our numerical example demonstrates that the proposed attack strategy is much more effective than the strategy proposed by an interesting related work [12] that studies how to add one adversarial data sample to influence the subspace estimated by PCA. Our study reveals that subspace learning via PCA is very sensitive to adversarial attacks. It is important to design adversarially robust subspace learning algorithms, which will be the focus of our future research.

2. PROBLEM FORMULATION

In this section, we introduce the problem formulation. Given a data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$ with each $\mathbf{x}_i \in \mathbb{R}^d$, our goal is to learn a low dimension subspace via PCA. In this paper, we consider an adversary setup in which an adversary will modify the data matrix $\mathbf{X}$ to $\hat{\mathbf{X}} = \mathbf{X} + \mathbf{\Delta X}$. Let $\mathbb{X}$ be a k-dimensional subspace learned from $\mathbf{X}$ and $\mathbb{Z}$ be a k-dimensional subspace learned from the modified data $\hat{\mathbf{X}}$. The goal of the adversary is to design the modification matrix $\mathbf{\Delta X}$ so as to make the distance between $\mathbb{X}$ and $\mathbb{Z}$ as large as possible. To measure such distance, we use the largest principal angle between $\mathbb{X}$ and $\mathbb{Z}$ as defined below [15].

Definition 1. *Let $\mathbb{X}$ and $\mathbb{Z}$ be two k-dimensional subspaces in $\mathbb{R}^d$, the principal angles $\{\theta_i\}_{i=1}^k$ are defined recursively:*

$$\begin{aligned}\cos(\theta_i) = \max_{\mathbf{u}_i\in\mathbb{X},\mathbf{v}_i\in\mathbb{Z}} \quad & \mathbf{u}_i^\top \mathbf{v}_i \\ s.t. \quad & \|\mathbf{u}_i\| = \|\mathbf{v}_i\| = 1, \\ & \mathbf{u}_j^\top \mathbf{u}_i = \mathbf{v}_j^\top \mathbf{v}_i = 0, \forall\, j = 1, 2, \cdots, i-1.\end{aligned}$$

It is easy to see $0 \le \theta_1 \le \theta_2 \le \cdots \le \theta_k \le \pi/2$. Here, we use the largest principal angle θ_k as the distance between two subspaces. This distance is also called Asimov distance. In the following, we will use $\theta(\mathbf{X}, \hat{\mathbf{X}})$ or simply θ to denote the Asimov distance between the subspace $\mathbb{X}$ estimated from $\mathbf{X}$ and the subspace $\mathbb{Z}$ estimated from $\hat{\mathbf{X}}$. Given orthonormal bases $\mathbf{U}_\mathbb{X}$ of $\mathbb{X}$ and orthonormal bases $\mathbf{U}_\mathbb{Z}$ of $\mathbb{Z}$, $\{\cos(\theta_1), \cos(\theta_2), \cdots, \cos(\theta_k)\}$ are the singular values of $\mathbf{U}_\mathbb{X}^\top \mathbf{U}_\mathbb{Z}$ [15]. Hence, the Asimov distance is determined by the smallest singular value of $\mathbf{U}_\mathbb{X}^\top \mathbf{U}_\mathbb{Z}$.

It is easy to see that, if no constraint is imposed on $\mathbf{\Delta X}$, then $\hat{\mathbf{X}}$ can be arbitrary and hence θ can be easily made to be $\pi/2$. In this paper, to make an initial attempt to understand the adversarial robustness of subspace learning algorithms, we impose energy and rank-one constraints on $\mathbf{\Delta X}$. In particular, we assume that the attack matrix $\mathbf{\Delta X}$ is a rank one matrix, and the energy of $\mathbf{\Delta X}$ is less than or equal to η. In this paper, we use the Frobenius norm $\|\mathbf{\Delta X}\|_\mathrm{F}$ to measure the energy. We note that the rank-one constraint is already powerful enough to capture the common modifications, for example, modifying one data sample, inserting one adversarial data, deleting one feature etc.

Formally, the optimal attack matrix that maximizes the Asimov distance between the subspace estimated from the original data set $\mathbf{X}$ and that from the modified data set $\hat{\mathbf{X}}$ under the constraints mentioned above can be written as

$$\begin{aligned}\max_{\mathbf{a}\in\mathbb{R}^d,\mathbf{b}\in\mathbb{R}^n} : \quad & \theta(\mathbf{X}, \hat{\mathbf{X}}) && (1)\\ \text{s.t.} \quad & \hat{\mathbf{X}} = \mathbf{X} + \mathbf{\Delta X}, \\ & \mathbf{\Delta X} = \mathbf{a}\mathbf{b}^\top, && (2)\\ & \|\mathbf{\Delta X}\|_\mathrm{F} \le \eta. && (3)\end{aligned}$$

Here, (2) is the rank-one constraint, and (3) is the energy constraint. It is easy to see that, for any feasible solution $(\tilde{\mathbf{a}}, \tilde{\mathbf{b}})$ with $||\tilde{\mathbf{b}}|| \neq 1$, we can construct another feasible solution $(||\tilde{\mathbf{b}}||\tilde{\mathbf{a}}, \tilde{\mathbf{b}}/||\tilde{\mathbf{b}}||)$ that gives the same objective function value. Hence, without loss of optimality, we will fix the norm of $\mathbf{b}$ to be 1 throughout of the paper.

3. OPTIMAL ADVERSARIAL STRATEGY ANALYSIS

In this section, we characterize the optimal solution to (1). We will first present our solution for the case where $\mathbf{X}$ has full column rank, and then generalize the result to the case where $\mathbf{X}$ does not necessarily have full column rank.

3.1. Full-Rank Case

In the full column rank case, $\mathrm{rank}(\mathbf{X}) = n$, where $n < d$. This case arises when the number of samples is limited, for example, at the beginning of online PCA. In the following, we first find the expression of $\theta(\mathbf{X}, \hat{\mathbf{X}})$ for any given $\hat{\mathbf{X}} = \mathbf{X} + \mathbf{a}\mathbf{b}^T$. Using this expression, we then characterize the optimal attack matrix $\mathbf{\Delta X}$.

Suppose the compact SVD of $\mathbf{X}$ is $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top = \mathbf{U}\mathbf{W}$, where $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_n)$. So, one set of orthonormal bases for the column space of $\mathbf{X}$ is $\mathbf{U}$. We can also use SVD to find a set of orthonormal bases $\tilde{\mathbf{U}}$ of $\mathrm{span}(\hat{\mathbf{X}})$.

Since $\hat{\mathbf{X}} = \mathbf{X} + \mathbf{a}\mathbf{b}^\top$ and according to [16], $\tilde{\mathbf{U}}$ can be directly expressed as a function of $\mathbf{U}$:

$$\tilde{\mathbf{U}} = \mathbf{U} + (\alpha\mathbf{U}\mathbf{w} + \beta\mathbf{s})\mathbf{w}^\top,$$

where

$$\begin{aligned}&\mathbf{a}_{u^\perp} = (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\mathbf{a},\ \mathbf{s} = \frac{\mathbf{a}_{u^\perp}}{\|\mathbf{a}_{u^\perp}\|},\ \tilde{\mathbf{w}} = -\mathbf{W}^{-\top}\mathbf{b},\\ &\mathbf{w} = \frac{\tilde{\mathbf{w}}}{\|\tilde{\mathbf{w}}\|},\ \omega = \frac{1}{\|\mathbf{a}_{u^\perp}\|}(1 - \mathbf{a}^\top\mathbf{U}\tilde{\mathbf{w}}),\ \mathbf{g} = [\tilde{\mathbf{w}}, \omega]^\top,\\ &\alpha = \frac{|\omega|}{\|\mathbf{g}\|} - 1,\ \beta = -\mathrm{sign}(\omega)\frac{\|\tilde{\mathbf{w}}\|}{\|\mathbf{g}\|}.\end{aligned}$$

Here $\mathbf{W}^{-\top} = (\mathbf{W}^{-1})^\top$. With this closed form expression for $\tilde{\mathbf{U}}$, we have

$$\mathbf{U}^\top\tilde{\mathbf{U}} = \mathbf{U}^\top(\mathbf{U} + (\alpha\mathbf{U}\mathbf{w} + \beta\mathbf{s})\mathbf{w}^\top) = \mathbf{I} + \alpha\mathbf{w}\mathbf{w}^\top.$$

The singular values of $\mathbf{I}+\alpha\mathbf{w}\mathbf{w}^\top$ are $\{1, 1, \cdots, 1+\alpha\mathbf{w}^\top\mathbf{w}\}$. Since $\mathbf{w}^\top\mathbf{w} = 1, 1 + \alpha = \frac{|\omega|}{\|\mathbf{g}\|}$, the smallest singular value of $\mathbf{U}^\top\tilde{\mathbf{U}}$ is $\cos(\theta) = \frac{|\omega|}{\|\mathbf{g}\|}$. Our objective is to maximize θ, which is equivalent to minimize the smallest singular value of $\mathbf{U}^\top\tilde{\mathbf{U}}$. Hence, the optimization problem (1) is simplified to

$$\begin{aligned}\min_{\mathbf{a},\mathbf{b}} : \quad & \frac{|\omega|}{\|\mathbf{g}\|} \\ \text{s.t.} \quad & \|\mathbf{a}\mathbf{b}^\top\|_\mathrm{F} = \|\mathbf{a}\|_2\|\mathbf{b}\|_2 \le \eta,\end{aligned}$$

where we use the identity $\|\mathbf{a}\|_2\|\mathbf{b}\|_2 = \|\mathbf{a}\mathbf{b}^\top\|_\mathrm{F}$. Expanding the objective function, we have

$$\frac{|\omega|}{\|\mathbf{g}\|} = \frac{|1 + \mathbf{a}_u^\top\mathbf{W}^{-\top}\mathbf{b}|}{\|[\|\mathbf{a}_{u^\perp}\|\mathbf{W}^{-\top}\mathbf{b}, 1 + \mathbf{a}_u^\top\mathbf{W}^{-\top}\mathbf{b}]\|}, \tag{4}$$

where $\mathbf{a}_u = \mathbf{U}^\top\mathbf{a}$.

Since $\mathbf{W} = \mathbf{\Sigma}\mathbf{V}^\top$, we have $\mathbf{W}^{-\top}\mathbf{b} = \mathbf{\Sigma}^{-1}\mathbf{V}^\top\mathbf{b}$. As $\mathbf{V}$ is an unitary matrix, changing the coordinate $\mathbf{b} \Leftarrow \mathbf{V}^\top\mathbf{b}$ does not change the constraint. So the value $\mathbf{a}_u^\top\mathbf{W}^{-\top}\mathbf{b}$ in the original coordinate is the same as $\mathbf{a}_u^\top\mathbf{\Sigma}^{-1}\mathbf{b}$ in the new coordinate. In the following, we will use this new coordinate system and hence the cost function in (4) can be written as

$$\frac{|\omega|}{\|\mathbf{g}\|} = \frac{|1 + \mathbf{a}_u^\top\mathbf{\Sigma}^{-1}\mathbf{b}|}{\|[\|\mathbf{a}_{u^\perp}\|\mathbf{\Sigma}^{-1}\mathbf{b}, 1 + \mathbf{a}_u^\top\mathbf{\Sigma}^{-1}\mathbf{b}]\|}. \tag{5}$$

The objective function (5) is zero if and only if the numerator is zero. Using the matrix norm inequality, we have

$$|\mathbf{a}_u^\top \mathbf{\Sigma}^{-1}\mathbf{b}| \le \|\mathbf{a}_u\|_2\|\mathbf{b}\|_2\|\mathbf{\Sigma}^{-1}\|_2 = \frac{1}{\sigma_n}\|\mathbf{a}_u\|_2\|\mathbf{b}\|_2 \overset{(a)}{\le} \frac{1}{\sigma_n}\|\mathbf{a}\|_2\|\mathbf{b}\|_2 = \frac{1}{\sigma_n}\|\mathbf{a}\mathbf{b}^\top\|_{\mathrm{F}} \overset{(b)}{\le} \frac{\eta}{\sigma_n},$$

where in (a) we use $\|\mathbf{a}_u\|_2 \le \|\mathbf{a}\|_2$, and (b) is due to the energy constraint. From the inequalities we conclude that when $\eta < \sigma_n$, we can not make the numerator to be zero. We now consider two different cases depending on whether we can make the numerator to be zero or not.

Case 1: When $\eta > \sigma_n$, if we set

$$\mathbf{a}_u = [0, 0, \cdots, -\sigma_n]^\top, \mathbf{b} = [0, 0, \cdots, 1]^\top,$$

and any $\|\mathbf{a}_{u^\perp}\|_2^2 = \hat{a}^2$ with $0 < \hat{a}^2 < \eta^2 - \sigma_n^2$, the numerator will be zero. Since $\mathbf{a} = \mathbf{U}\mathbf{a}_u + (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\mathbf{a}_{u^\perp}$, the attacker can make the Asimov distance to be $\pi/2$ by setting:

$$\mathbf{a} = -\sigma_n \mathbf{u}_n + \hat{a}\mathbf{u}_q, \mathbf{b} = \mathbf{v}_n, \tag{6}$$

where $\mathbf{u}_q$ is any vector orthogonal to the column space of $\mathbf{X}$ and $0 < \hat{a}^2 < \eta^2 - \sigma_n^2$.

Case 2: When $\eta \le \sigma_n$, the value of $1 + \mathbf{a}_u^\top \mathbf{\Sigma}^{-1}\mathbf{b}$ can not reach zero. In this case, it is easy to check that minimizing (5) is equivalent to maximizing

$$\frac{\|\mathbf{a}_{u^\perp}\|_2^2 \|\mathbf{\Sigma}^{-1}\mathbf{b}\|_2^2}{(1 + \mathbf{a}_u^\top \mathbf{\Sigma}^{-1}\mathbf{b})^2}. \tag{7}$$

With the norm of $\mathbf{b}$ being 1, $\|\mathbf{\Sigma}^{-1}\mathbf{b}\|_2^2$ is maximized when $\mathbf{b} = [0, 0, \cdots, 1]^\top$. Furthermore, for any fixed norm of $\mathbf{a}_u$, $(1+\mathbf{a}_u^\top \mathbf{\Sigma}^{-1}\mathbf{b})^2$ is minimized when $\mathbf{a}_u = [0, 0, \cdots, -\|\mathbf{a}_u\|_2]^\top$, $\mathbf{b} = [0, 0, \cdots, 1]^\top$. Hence, for any fixed norms of $\mathbf{a}_u$, $\mathbf{a}_{u^\perp}$, the objective function (7) is maximized when

$$\mathbf{a}_u = [0, 0, \cdots, -\|\mathbf{a}_u\|_2]^\top, \quad \mathbf{b} = [0, 0, \cdots, 1]^\top. \tag{8}$$

Let $c = \|\mathbf{a}_{u^\perp}\|_2, h = \|\mathbf{a}_u\|_2$, and use the optimal form of $\mathbf{a}_u$ and $\mathbf{b}$ in (8), the objective function (7) can be simplified to

$$\begin{aligned} \max_{c,h}: \quad & \frac{c^2/\sigma_n^2}{(1 - h/\sigma_n)^2} \\ \text{s.t.} \quad & (c^2 + h^2) \le \eta^2, \end{aligned} \tag{9}$$

It is easy to check that the objective function is maximized when $c^2 + h^2 = \eta^2$. Hence, we have $c^2 = \eta^2 - h^2$. Inserting this value of c into the objective function and setting the derivative with respect to h to be 0, we get a unique solution $h = \eta^2/\sigma_n$. At this value of h, the second derivative is $\frac{-2\sigma_n^2}{(\sigma_n^2 - \eta^2)^3}$, which is negative. It indicates that $h = \eta^2/\sigma_n$ is indeed the maximum point. Hence $c = \pm\eta\sqrt{1 - \eta^2/\sigma_n^2}$. This implies that the optimal solution of problem (1) for Case 2 is

$$\mathbf{a} = -\eta^2/\sigma_n \mathbf{u}_n \pm \eta\sqrt{1 - \eta^2/\sigma_n^2}\mathbf{u}_q, \mathbf{b} = \mathbf{v}_n. \tag{10}$$

Combining Cases 1 and 2, we have that the optimal value of problem (1) in the full-rank case is

$$\theta^* = \begin{cases} \pi/2, & \text{if } \eta > \sigma_n \\ \arcsin(\eta/\sigma_n), & \text{if } \eta \le \sigma_n \end{cases}. \tag{11}$$

3.2. Low-Rank Case

We now consider the case where $\mathbf{X}$ is not full rank. Let $k < \min(d, n)$ be the rank of $\mathbf{X}$, and denote function $g_k(\cdot)$ as the PCA operation that computes the k leading principal components. In this subsection, with a slight abuse of notation, we write the full SVD of $\mathbf{X}$ as $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. The optimal attack matrix could be found by solving

$$\begin{aligned} \max_{\mathbf{a}\in\mathbb{R}^d, \mathbf{b}\in\mathbb{R}^n}: \quad & \theta(\mathbf{X}, g_k(\hat{\mathbf{X}})) \\ \text{s.t.} \quad & \hat{\mathbf{X}} = \mathbf{X} + \mathbf{a}\mathbf{b}^\top, \\ & \|\mathbf{a}\|_2\|\mathbf{b}\|_2 \le \eta. \end{aligned} \tag{12}$$

We can further simplify this optimization problem as

$$\begin{aligned} \max_{\mathbf{a}\in\mathbb{R}^{k+1}, \mathbf{b}\in\mathbb{R}^{k+1}}: \quad & \theta(\tilde{\mathbf{\Sigma}}, g_k(\mathbf{Y})) \\ \text{s.t.} \quad & \mathbf{Y} = \tilde{\mathbf{\Sigma}} + \mathbf{a}\mathbf{b}^\top, \\ & \|\mathbf{a}\|_2\|\mathbf{b}\|_2 \le \eta, \end{aligned} \tag{13}$$

where $\tilde{\mathbf{\Sigma}} = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_k, 0)$ and $\{\sigma_1, \sigma_2, \cdots, \sigma_k\}$ are singular values of $\mathbf{X}$. Due to the space limitation, we omit the detailed proof of the equivalence between (12) and (13) and only provide the main idea of the proof here. The main idea of the simplification is to left multiply the unitary matrix $\mathbf{U}^\top$ and right multiply the unitary matrix $\mathbf{V}$ on both $\mathbf{X}$ and $\hat{\mathbf{X}}$. Note that multiplying a unitary matrix does not change the column space and its singular values. In addition, a rank-one modification can only add at most one principal component orthogonal to its original column subspace. Hence, by changing the coordinates, $\mathbf{a}$ and $\mathbf{b}$ are $(k+1)$-dimensional vectors.

When $\eta > \sigma_k$, it is simple to verify that the simple solution $\mathbf{a} = [0, 0, \cdots, \eta]^\top$, $\mathbf{b} = [0, 0, \cdots, 1, 0]^\top$ leads to the maximal Asimov distance, which is $\pi/2$.

When $\eta \le \sigma_k$, the following theorem characterizes the form of optimal $\mathbf{a}$ and $\mathbf{b}$.

Theorem 1. *There exists an optimal solution of problem* (13) *in the following form*

$$\mathbf{a} = [0, \cdots, 0, a_k, a_{k+1}]^\top, \mathbf{b} = [0, 0, \cdots, 0, 1, 0]^\top, \tag{14}$$

with $a_k^2 + a_{k+1}^2 = \eta^2$.

Proof. Due to space limitation, we omit the proof details. □

Since $\|\mathbf{a}\|_2^2 = \eta^2$ and $\mathbf{a}$ is in the form of (14), we can write $\mathbf{a} = \eta[0, 0, \cdots, \cos(\alpha), \sin(\alpha)]^\top$, where $\alpha \in [0, 2\pi)$.

To compute the k leading principal components of $\mathbf{Y}$, we can perform the eigenvalue decomposition of $\mathbf{Y}\mathbf{Y}^\top$,

$$\mathbf{Y}\mathbf{Y}^\top = \begin{bmatrix} \mathbf{\Lambda}_{k-1}^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{c}\mathbf{c}^\top \end{bmatrix}, \tag{15}$$

where $\mathbf{c} = [\sigma_k + \eta\cos\alpha, \eta\sin(\alpha)]^\top$, $\mathbf{\Lambda}_{k-1} = \mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_{k-1})$. Suppose the compact SVD of $\mathbf{Y}\mathbf{Y}^\top$ is $\mathbf{Y}\mathbf{Y}^\top = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^\top$, where

$$\hat{\mathbf{U}} = \begin{bmatrix} \mathbf{I}_{k-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{z} \end{bmatrix}, \tag{16}$$

and $\mathbf{z} \in \mathbb{R}^2$ is the eigenvector of $\mathbf{c}\mathbf{c}^\top$ corresponding to its nonzero eigenvalue. Since one set of orthonormal bases of $\mathrm{span}(\tilde{\mathbf{\Sigma}})$ is $[\mathbf{I}_k, \mathbf{0}]^\top$, the Asimov distance is determined by the singular values of

$$\begin{bmatrix} \mathbf{I}_k \\ \mathbf{0} \end{bmatrix}^\top \cdot \begin{bmatrix} \mathbf{I}_{k-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{z} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{k-1} & \mathbf{0} \\ \mathbf{0} & z_1 \end{bmatrix}.$$

So, the Asimov distance is $\arccos(|z_1|)$. Since $\mathbf{c}$ is the eigenvector of $\mathbf{c}\mathbf{c}^\top$ corresponding to its nonzero eigenvalue, then $|z_1| = \frac{|c_1|}{\|\mathbf{c}\|}$. Our objective function reduces to

$$\min_{\alpha\in[0,2\pi)} : \quad \frac{|\sigma_k + \eta\cos(\alpha)|}{\|[\sigma_k + \eta\cos(\alpha), \eta\sin(\alpha)]\|_2}.$$

For this problem, we can show that the optimal α is $\alpha = \arccos(-\eta/\sigma_k)$ or $2\pi - \arccos(-\eta/\sigma_k)$. Hence, the optimal solution of problem (13) is

$$\mathbf{a} = \left[0, 0, \cdots, -\eta^2/\sigma_k, \pm\eta\sqrt{1-\eta^2/\sigma_k^2}\right]^\top, \tag{17}$$

$$\mathbf{b} = [0, 0, \cdots, 0, 1, 0]^\top, \tag{18}$$

which indicates the optimal solution of problem (12) is

$$\mathbf{a} = -\eta^2/\sigma_k \mathbf{u}_k \pm \eta\sqrt{1-\eta^2/\sigma_k^2}\mathbf{u}_q, \mathbf{b} = \mathbf{v}_k, \tag{19}$$

where $\mathbf{u}_q$ is any vector orthogonal to the column space of $\mathbf{X}$. The corresponding optimal subspace distance is $\theta^* = \arcsin(\eta/\sigma_k)$. In summary, the optimal Asimov distance in the low-rank case is:

$$\theta^* = \begin{cases} \pi/2, & \text{if } \eta > \sigma_k \\ \arcsin(\eta/\sigma_k), & \text{if } \eta \le \sigma_k \end{cases}, \tag{20}$$

which is similar to the full column rank case. In conclusion, the optimal subspace distance only depends on the smallest singular value and the adversary's energy budget in both full column rank and low-rank cases.

4. NUMERICAL EXAMPLE

In this section, we provide a numerical example to illustrate the results obtained in this paper. In our simulation, we set $d = 5$, $n = 5$, and $k = 3$. We generate the original data as $\mathbf{X} = \mathbf{A}\mathbf{B}^\top$, where $\mathbf{A} \in \mathbb{R}^{d\times k}, \mathbf{B} \in \mathbb{R}^{n\times k}$ and each entry of $\mathbf{A}, \mathbf{B}$ is i.i.d. generated according to the standard normal distribution. First, we use our optimal attack strategy to design $\mathbf{a}, \mathbf{b}$ and add the attack matrix $\mathbf{\Delta X} = \mathbf{a}\mathbf{b}^\top$ to the original data matrix $\mathbf{X}$. We then perform SVD on $\hat{\mathbf{X}}$ and select the k leading principal components. Finally, we compute the distance between the selected subspace and the original subspace. In the simulation, we also compare the strategy described in [12], which adds one adversarial data sample into the data set. We denote this strategy as adPCA. In addition, we also conduct a test using a random attack strategy, in which we randomly generate $\mathbf{a}, \mathbf{b}$ with each entry of $\mathbf{a}, \mathbf{b}$ being i.i.d. generated according to the standard normal distribution. Then we normalize the energy of $\mathbf{a}\mathbf{b}^\top$ to be η^2. For each η, we repeatedly generate 100000 pairs of $\mathbf{a}$ and $\mathbf{b}$. For each pair of $\mathbf{a}$ and $\mathbf{b}$, we compute its corresponding Asimov distance.

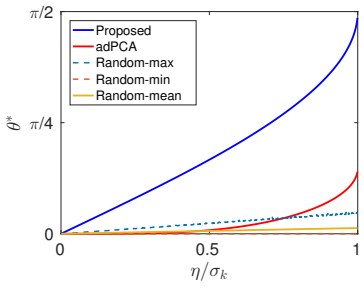

Fig. 1. Subspace distance with different attack strategies under different energy ratios.

Fig.1 illustrates the Asimov distances computed by these three different strategies. In this figure, the x axis is the ratio between η and the smallest singular value of $\mathbf{X}$. For the random strategy, we plot the mean, maximal, and minimal subspace distance for each η/σ_k. From the figure, we can see that our strategy is better than adPCA. This is due to the fact that our method can modify all the data samples, while adPCA can only add one adversarial data sample into the data set. Hence, our strategy has more degrees of freedom to manipulate the data samples. From the figure, we can see that our optimal attack strategy can achieve a significant larger Asimov distance than that can be achieved by the random attack strategy.

5. CONCLUSION

In this paper, we have investigated the adversarial robustness of PCA problem. We have characterized the optimal rank-one adversarial modification strategy for the attacker to modify the data. The strategy only depends on the smallest singular value of the original data matrix and the adversary's energy budget. In the future, it is of interest to investigate the defense strategy to mitigate the effects of this attack.

6. REFERENCES

[1] H. Guo, C. Qiu, and N. Vaswani, "An online algorithm for separating sparse and low-dimensional signal sequences from their sum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4284–4297, Aug. 2014.

[2] R. Otazo, E. J. Candès, and D. K. Sodickson, "Low-rank plus sparse matrix decomposition for accelerated dynamic mri with separation of background and dynamic components," *Magnetic Resonance in Medicine*, vol. 73, no. 3, pp. 1125–1136, Apr. 2015.

[3] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009.

[4] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *Journal of the ACM*, vol. 58, no. 3, pp. 11:1–11:37, June 2011.

[5] B. Huval, T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, P. Rajpurkar, T. Migimatsu, R. Cheng-Yue, et al., "An empirical evaluation of deep learning on highway driving," *arXiv preprint arXiv:1504.01716*, Apr. 2015.

[6] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Communications Surveys Tutorials*, vol. 18, no. 2, pp. 1153–1176, Oct. 2016.

[7] G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-Van De Kaa, P. Bult, B. Van Ginneken, and J. Van Der Laak, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," *Scientific reports*, vol. 6, no. 1, pp. 26286, May 2016.

[8] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, Dec. 2014.

[9] B. Biggio, I. Pillai, S. Rota Bulò, D. Ariu, M. Pelillo, and F. Roli, "Is data clustering in adversarial settings secure?," in *Proc. ACM Workshop on Artificial Intelligence and Security*, Berlin, Germany, Nov. 2013, pp. 87–98.

[10] S. Alfeld, X. Zhu, and P. Barford, "Data poisoning attacks against autoregressive models," in *Proc. AAAI Conference on Artificial Intelligence*, Phoenix, Arizona, Feb. 2016, pp. 1452–1458.

[11] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?," in *Proc. ACM Symposium on Information, Computer and Communications Security*, Taipei, Taiwan, Mar. 2006, pp. 16–25.

[12] D. L. Pimentel-Alarcn, A. Biswas, and C. R. Sols-Lemus, "Adversarial principal component analysis," in *IEEE International Symposium on Information Theory*, Aachen, Germany, June 2017, pp. 2363–2367.

[13] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, Dec. 2013.

[14] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, Massachusetts, June 2015, pp. 427–436.

[15] G. H. Golub and C. F. Van Loan, *Matrix computations*, The Johns Hopkins University Press, 2013.

[16] R. Zimmermann, "A closed-form update for orthogonal matrix decompositions under arbitrary rank-one modifications," *arXiv preprint arXiv:1711.08235*, Nov. 2017.